Practical Small Sample Asymptotics for Regression Problems

Robert L. Strawderman *
University of Michigan

George Casella †and Martin T. Wells ‡
Cornell University

December 20, 1994

Abstract

Saddlepoint approximations are derived for sums of independent, but not necessarily identically distributed, random variables, along with generalizations to estimating equations and multivariate problems. These results are exactly what is needed to obtain accurate approximations to the distributions of regression coefficients. We give general formulae for the distribution of the coefficients in a generalized linear model with both canonical and non-canonical link functions, and illustrate the accuracy of our approximations with examples of exponential, Poisson, and logistic regression. Finally, we apply our approximations to an actual data set, and show how the Gibbs sampler may be used to obtain confidence sets for each regression parameter based on the saddlepoint approximation to the joint density.

Key words: Estimating Equations; Generalized Linear Model; Gibbs Sampling; Laplace's Method; Logistic Regression; Maximum Likelihood Estimation; Monte Carlo Integration

---

*Part of this research was performed while Prof. Strawderman was visiting Cornell University, supported by National Science Foundation Grant No. DMS-9305547

†Research supported by National Science Foundation Grant No. DMS-9305547. This is paper BU-1269-M in the Biometrics Unit, Cornell University, Ithaca NY 14850

‡Research supported by NIH Grant No. RO1-CA61120.AMS

1 Introduction

Daniels (1954) seminal paper on saddlepoint approximations in statistics spawned a great deal of research in the general area of small sample asymptotic approximations. His original paper concentrates primarily upon the distribution of the sample mean in the case where the random variables of interest are independent and identically distributed. With few exceptions, publications in this area have relied upon this assumption as well, ostensibly ruling out many important applications. In particular, Barndorff-Nielsen and Cox (1979) derive saddlepoint approximations for independent and identically distributed multivariate random variables, and then discuss methods for obtaining saddlepoint approximations to conditional distributions by various means. Their "double saddlepoint" approximation requires the density of the random variables being conditioned upon to be approximated as well. Such an approach may not be particularly advantageous in regression problems, for example, where one is forced to assume a parametric form for the distribution of the predictors. Durbin (1980) derives density approximations for sufficient statistics without assuming that the random variables are independent or identically distributed by using arguments similar to Daniels (1954). An interesting aspect of Durbin's approach is that the calculation of the saddlepoint is bypassed by choosing the contour of integration to pass through the observed point (i.e. the point at which the density approximation is desired). Skovgaard (1987) derives saddlepoint approximations for conditional distributions of the form $P(\overline{Y} \geq y | A'\overline{X} = a)$, where A is a matrix with certain properties and $(\overline{X}, \overline{Y})$ is the bivariate sample mean. Davison (1988) derives formulae for approximating conditional distributions in generalized linear models, but via an approach similar to the "double-saddlepoint" approximation of Barndorff-Nielsen and Cox (1979).

A goal of this paper is to extend the results of Daniels (1954) to sums of independent, but not necessarily identically distributed, random variables. Details of the derivation are given in the Appendix; for simplicity, we deal with continuous random variables only, the case of

lattice random variables being entirely similar and therefore omitted. In Section 2, we extend the results of Daniels (1983), deriving formulae for saddlepoint approximations to the distribution of univariate maximum likelihood estimators. Extensions to multivariate case are given in Section 2.2, and rely upon the connection between the results of Daniels (1983) and Field (1982). Approximations within the class of generalized linear models are particularly straightforward in this setting; these are discussed in Section 3.1. Some simulated examples are provided in Sections 3.2 to demonstrate the accuracy of the approximation in various situations. We close the paper in Section 4 with an implementation of our methodology to an actual data set, one previously analyzed by Härdle and Stoker (1989).

The crucial conditions for these asymptotic expansions to be valid (in either case) appear to be similar to the conditions under which this methodology was originally developed. The presentation throughout the paper is somewhat informal, and contains little discussion surrounding the precise technical conditions under which these approximations hold; we leave this for future work. However, we have tried to be very explicit in our notation in the hope that this helps to clarify some of the details of the saddlepoint approximations.

2 Saddlepoint Expansions for Estimating Equations

The problem of approximating the small sample distribution of estimators derived from estimating equations, such as (1) below, has already received much attention. In particular, approximations for general M-estimators has been studied in depth for independent and identically distributed random variables; see, for example, Field (1982), Field and Hampel (1982), Daniels (1983), and Ronchetti and Welsh (1994). However, such results do not extend to the regression setting unless (i) the errors are additive and (ii) the distribution of the covariates being conditioned upon is known or can be approximated. Saddlepoint methods can be extended to situations where the statistic of interest can be expressed as (or is related to) a sum of independent but not necessarily identically distributed random

variables; the details are provided in the Appendix. The intent of this section is to show how these results may be applied in approximating the distribution of estimators derived from estimating equations; the proof in the univariate case follows that of Daniels (1983), and results for the multivariate case may be obtained by noting some connections between the results of Field (1982), Daniels (1983), and Ronchetti and Welsh (1994).

Let $X_1, \ldots, X_n$ be n independent random vectors. Suppose that X_i has density $f_i(\cdot|\theta)$, where the support of $f_i(\cdot|\theta)$ is some subset of $\mathbb{R}^p$, $\theta \in \Theta$ where $\Theta \subseteq \mathbb{R}^k$ for $k \geq 1$, and the subscript i on the density $f_i(\cdot|\theta)$ allows for dependence upon a vector of covariates. Let $\psi_i(\cdot|\theta) = \partial \log f_i(\cdot|\theta)/\partial\theta$; suppose that $E[\psi_i(X_i|\theta_0)] = 0, i = 1 \ldots n$ whenever θ_0 is the true parameter. Define

$$\overline{W}(a) = \frac{1}{n}\sum_{i=1}^{n} \psi_i(X_i|a), \tag{1}$$

where $a \in \Theta$, and let $\hat{\theta}$ solve $\overline{W}(\hat{\theta}) = 0$. We assume that a unique solution to this set of k equations exists on the interior of Θ. In addition, we assume that the cumulant generating function for $n\overline{W}(a)$, say $K_n(t|a, \theta_0)$, exists for $t \in B(\varepsilon)$ where $B(\varepsilon) \subseteq \mathbb{R}^k$ is an open ball of radius $\varepsilon > 0$ about $t = 0$. The notation emphasizes the fact that the saddlepoint depends both upon the point in question (a) and also on the true parameter value (θ_0). The case where θ is a scalar parameter is treated in the next section; the extension to vector parameters is discussed in Section 2.2.

2.1 Scalar Parameters

If $\overline{W}(a)$ is a monotone decreasing function in a having a unique solution in an open neighborhood about θ_0 with probability going to one, then the necessary assumptions of existence, etc ...in the case of a scalar parameter θ are implicitly retained in the well-known relation

$$\mathrm{pr}_{\theta_0}\{\hat{\theta} > a\} \equiv \mathrm{pr}_{\theta_0}\{\overline{W}(a) > 0\}$$

(Small and McLeish, 1994, page 87). Since $n\overline{W}(a)$ is a sum of independent random variables, the results derived in the Appendix apply to the right-hand side of this expression under the appropriate regularity conditions.

Specifically, from (17) we immediately obtain that

$$\begin{aligned} \mathrm{pr}_{\theta_0}\{\overline{W}(a) > 0\} &= \mathrm{pr}_{\theta_0}\{n\overline{W}(a) > 0\} \\ &= \int_{W_0}^{\infty} \left(\frac{nR_n''(t|a,\theta_0)}{2\pi}\right)^{1/2} \exp\left(n[R_n(t|a,\theta_0) - tR_n'(t|a,\theta_0)]\right)\, dt, \end{aligned}$$

where $R_n(t|a,\theta_0) = n^{-1}K_n(nt|a,\theta_0)$ and W_0 solves $R_n'(W_0|a,\theta_0) = 0$. This expression may be manipulated to obtain a saddlepoint approximation to the density of $\hat{\theta}$. Assuming that there exists a function $0 < C(a,\theta_0) < \infty$ such that

$$\lim_{t\to 0} \frac{\partial}{\partial\theta}\left(-\frac{R_n(t|a,\theta_0)}{t}\right) = C(a,\theta_0), \tag{2}$$

the saddlepoint approximation to the density of $\hat{\theta}$ is then given by

$$\begin{aligned} g_{\hat{\theta}}(a|\theta_0) &= \left(\frac{n}{2\pi R_n''(W_0|a,\theta_0)}\right)^{1/2} \left(-\frac{\partial}{\partial\theta}\frac{R_n(W_0|a,\theta_0)}{W_0}\right) \exp\left(nR_n(W_0|a,\theta_0)\right) \\ &= \left(\frac{1}{2\pi K_n''(nW_0|a,\theta_0)}\right)^{1/2} \left(-\frac{\partial}{\partial\theta}\frac{K_n(nW_0|a,\theta_0)}{nW_0}\right) \exp\left(K_n(nW_0|a,\theta_0)\right)\ , \end{aligned} \tag{3}$$

where the latter expression follows from the definition of $R_n(t|a,\theta_0)$.

If primary interest lies in calculating tail probabilities, one could use the analogous form to the Lugannani and Rice (1980) formula. The appropriate version of the formula is

$$\mathrm{pr}_{\theta_0}\{\hat{\theta} > a\} \doteq 1 - \Phi(y) + \frac{\exp\left(K_n(nW_0|a,\theta_0)\right)}{\sqrt{2\pi}}\left(\frac{1}{z} - \frac{1}{y}\right), \tag{4}$$

where $y = \mathrm{sign}(W_0)\left(-2K_n(nW_0|a,\theta_0)\right)^{1/2}$ and $z = W_0\left(n^2K_n''(nW_0|a,\theta_0)\right)^{1/2}$. We note that both (3) and (4) reduce to the formulae of Daniels (1983) when the random variables $X_1,\ldots,X_n$ are independent and identically distributed.

2.2 Vector Parameters

Field (1982) derives approximations in the case of independent and identically distributed random variables. A careful inspection of his results demonstrates that his formulae are generalizations of those given in Daniels (1983). The key to their equivalence lies in the relationship described in (2). Specifically, in the case of independent and identically distributed random variables, Daniels (1983) remarks that the limit $C(a, \theta_0)$ is actually $\int \psi'(x|a) f(x|\theta_0) dx$, where $f(x|\theta_0)$ is the density of each random variable. The analogous formula for a vector parameter is given in Field (1982) and also Ronchetti and Welsh (1994) as

$$A_F(t|a, \theta_0) = \exp\left(-K(t|a, \theta_0)\right) \int \frac{\partial}{\partial a} \psi(x|a) \exp\left(t'\psi(x|a)\right) f(x|\theta_0)\, dx, \qquad (5)$$

where t is the $k \times 1$ tilting parameter and is chosen so that the random vector $\psi(X|a)$ has mean zero under the conjugate density.

The derivation in the case where the $X_i's$ are independent but not identically distributed is essentially the same, except for the fact that the moment generating function (and hence the cumulant generating function) is more complex. In this case, we choose the tilting parameter t so that the derivative of the cumulant generating function of $\sum_{i=1}^n \psi_i(X|a)$ equals zero, and hence so that the mean of the conjugate density equals zero. Under this normalization, the methods of Field (1982) may be extended to M-estimators for independent but not identically distributed random vectors. An appropriate generalization of the regularity conditions given in Field (1982) is also required. In the case of a curved exponential family, these regularity conditions can be weakened in the style of Hougaard (1985) without diminishing the strength of the results.

Let $K_n(t|a, \theta_0) = \sum_{i=1}^n K_i(t|a, \theta_0)$, where

$$K_i(t|a, \theta_0) = \log \int \exp\left(t'\psi_i(x|a)\right) f_i(x|\theta_0)\, dx.$$

As before, define $R_n(t|a,\theta_0) = n^{-1}K_n(nt|a,\theta_0)$; then,

$$R_n'(t|a,\theta_0) \;=\; \sum_{i=1}^{n} \exp\left(-K_i(nt|a,\theta_0)\right) \int \psi_i(x|a)\, \exp\left(nt'\psi_i(x|a)\right)\, f_i(x|\theta_0)\, dx$$

and

$$R_n''(t|a,\theta_0) \;=\; n\sum_{i=1}^{n} \exp\left(-K_i(nt|a,\theta_0)\right) \int \psi_i(x|a)\, \psi_i'(x|a)\, \exp\left(nt'\psi_i(x|a)\right)\, f_i(x|\theta_0)\, dx \quad (6)$$

A natural generalization of the matrix $A_F(t|a,\theta_0)$ defined in (5) is given by

$$A(t|a,\theta_0) \;=\; \sum_{i=1}^{n} \exp\left(-K_i(nt|a,\theta_0)\right) \int \frac{\partial}{\partial a}\psi_i(x|a)\, \exp\left(nt'\psi_i(x|a)\right)\, f_i(x|\theta_0)\, dx. \quad (7)$$

The tilt vector, denoted by W_0, solves $R_n'(W_0|a,\theta_0) = 0$ or equivalently $K_n'(nW_0|a,\theta_0) = 0$, and therefore the unnormalized approximation to the density of $\hat{\theta}$ is given by

$$\begin{aligned} g_{\hat{\theta}}(a|\theta_0) \;&=\; \left(\frac{n}{2\pi}\right)^{k/2} \exp\left(nR_n(W_0|a,\theta_0)\right)\, \left|R_n''(W_0|a,\theta_0)\right|^{-1/2} \left|A(W_0|a,\theta_0)\right| \\ &=\; \left(\frac{1}{2\pi}\right)^{k/2} \exp\left(K_n(nW_0|a,\theta_0)\right)\, \left|K_n''(nW_0|a,\theta_0)\right|^{-1/2} \left|A(W_0|a,\theta_0)\right|. \end{aligned}$$

The error in this approximation under favorable conditions should be $O(n^{-1})$; since $g_{\hat{\theta}}(\cdot|\theta_0)$ need not integrate to one, renormalizing by its integral provides a uniform error bound of $O(n^{-3/2})$ in the so-called normal deviation region. Ronchetti and Welsh (1994) remark that such approximations will be particularly good in situations where the estimating function is bounded; we expect that similar results hold here.

If we restrict our attention to the case of maximum likelihood estimation, then by noting that

$$\frac{\partial}{\partial a}\psi_i(x|a) \;=\; \psi_i(x|a)\, \psi_i'(x|a)$$

we get the simplified formula

$$g_{\hat{\theta}}(a|\theta_0) \;=\; \left(\frac{1}{2\pi}\right)^{k/2} \exp\left(K_n(nW_0|a,\theta_0)\right)\, \left|K_n''(nW_0|a,\theta_0)\right|^{1/2}. \quad (8)$$

As one might expect, this formula is exact in the case of linear regression with normal errors. Suppose that $Y = X\theta + \varepsilon$, where Y is $n \times 1$, X is $n \times k$, θ is $k \times 1$, and $\varepsilon \sim N(0, V)$ is

$n \times 1$. The distribution of the maximum likelihood estimator $\hat{\theta}$ is $N_k(\theta_0, (X'V^{-1}X)^{-1})$ when the elements of the V are assumed to be known and θ_0 is the true mean. The saddlepoint approximation requires the cumulant generating function for the score function; this is easily obtained from the normal equations, and yields

$$K_n(t|a, \theta_0) = t'X'V^{-1}X(\theta_0 - a) + t'X'V^{-1}Xt,$$

where t is $k \times 1$. The saddlepoint W_0, determined as the solution to $K'(nW_0|a, \theta_0) = 0$, is $n^{-1}(a - \theta_0)$. Returning to (8), we see that

$$\begin{aligned} g_{\hat{\theta}}(a|\theta_0) &= \left(\frac{1}{2\pi}\right)^{k/2} \exp\left(K_n(nW_0|a, \theta_0)\right) \left|K_n''(nW_0|a, \theta_0)\right|^{1/2} \\ &= \left(\frac{1}{2\pi}\right)^{k/2} \exp\left(-\frac{(a - \theta_0)'X'V^{-1}X(a - \theta_0)}{2}\right) \left|X'V^{-1}X\right|^{1/2}, \end{aligned}$$

which is exactly the density for a $N_k(\theta_0, (X'V^{-1}X)^{-1})$ random vector evaluated at the point a, which is what we desired to show. In the Sections 3.2-3.4, we present some examples where the approximation, although not exact, is extremely accurate.

3 Applications to Generalized Linear Models

3.1 General Theory

Let $Y_1, \ldots, Y_n$ be independent random variables with respective means $\mu_i(\theta)$, where $g(\mu_i(\theta)) = \alpha_i + \theta' z_i$ for some monotonic differentiable function $g(\cdot)$, α_i are known constants, and z_i is a $k \times 1$ vector of covariates. Suppose that Y_i has the density or mass function

$$f_i(y|\theta) = \exp\left(\frac{y\,\eta_i - c(\eta_i)}{\phi_i} + d_i(y, \phi_i)\right),$$

where $\eta_i = \eta_i(\theta)$ and $\dot{c}(\eta_i) = \partial c(\eta_i)/\partial \eta_i = \mu_i(\theta)$. This is the standard regression set-up for a generalized linear model with a non-canonical link, individual offset parameters, and known dispersion (McCullagh and Nelder, 1989). Let us further assume that $\phi_i = \phi w_i$ for known ϕ

and weights w_i. Then, the score function for the $k \times 1$ parameter vector θ may be written as

$$l'(\theta) \;=\; \sum_{i=1}^{n} z_i \left(Y_i - \mu_i(\theta)\right) \frac{b_i(\theta)}{\phi w_i} \;=\; \sum_{i=1}^{n} \psi_i(Y_i|\theta)$$

where $b_i^{-1}(\theta) = V(\mu_i(\theta))\ \dot{g}(\mu_i(\theta))$, $V(\mu_i(\theta))$ is the variance function associated with the family of densities, and $\mathrm{var}(Y_i) = \phi w_i V(\mu_i(\theta))$ (Hinkley *et al.* 1991, Chapter 4).

Elementary algebraic manipulations yield the cumulant generating function for $\sum_{i=1}^{n} \psi_i(Y_i|a)$ when the true parameter is θ_0; this may be expressed as

$$K_n(t|a,\theta_0) \;=\; \sum_{i=1}^{n} \left[c(\eta_i(\theta_0) + t' z_i b_i(a)) - c(\eta_i(\theta_0)) - \mu_i(a) b_i(a) t' z_i\right] (\phi w_i)^{-1}. \qquad (9)$$

Since $R_n(t|a,\theta_0) = n^{-1} K_n(nt|a,\theta_0)$, then by noting that

$$\frac{d}{dt} c(\eta_i(\theta_0) + nt' z_i b_i(a)) \;=\; n z_i b_i(a) g^{-1}(\eta_i(\theta_0) + nt' z_i b_i(a)),$$

it follows that

$$R_n'(t|a,\theta_0) \;=\; \sum_{i=1}^{n} z_i \frac{b_i(a)}{\phi w_i} \left[g^{-1}(\eta_i(\theta_0) + nt' z_i b_i(a)) - g^{-1}(\alpha_i + a' z_i)\right]. \qquad (10)$$

The saddlepoint W_0 is chosen to satisfy $R_n'(W_0|a,\theta_0) = 0$; under appropriate conditions on the $z_i's$ (e.g. not identically zero), the saddlepoint will be unique. However, it does not exist in closed form for a general link function, and thus for fixed θ_0 the saddlepoint needs to be determined for each value of a using numerical techniques. In order to compute the saddlepoint approximation to the density we need $K_n''(nW_0|a,\theta_0)$; it has a relatively simple form, and is readily computed once the saddlepoint is known. Specifically, using (6) we may show that

$$K_n''(nt|a,\theta_0) = n R_n''(t|a,\theta_0) \;=\; \sum_{i=1}^{n} z_i z_i' \frac{b_i^2(a)}{\phi^2 w_i^2}\ h_i(a,\theta_0,W_0),$$

where

$$h_i(a,\theta_0,W_0) \;=\; \int (y - \mu_i(a))^2\ \exp\left(\frac{y\,\hat{\gamma}_i - c(\hat{\gamma}_i)}{\phi w_i} + d_i(y,\phi_i)\right)\ dy$$

and $\hat{\gamma}_i = nW_0' z_i b_i(a) + \eta_i(\theta_0)$. Expanding out the square and using standard properties of exponential families, one may show that

$$K_n''(nt|a,\theta_0) = \sum_{i=1}^{n} z_i z_i' \frac{b_i^2(a)}{\phi^2 w_i^2} \left[w_i \phi \ddot{c}(\hat{\gamma}_i) + \dot{c}^2(\hat{\gamma}_i) - 2\mu_i(a)\dot{c}(\hat{\gamma}_i) + \mu_i^2(a) \right]; \quad (11)$$

note that the term in the brackets has the interpretation of variance plus bias squared. The saddlepoint approximation to the density of $\hat{\theta}$ is then given by

$$g_{\hat{\theta}}(a|\theta_0) = \left(\frac{1}{2\pi}\right)^{k/2} \exp\left(K_n(nW_0|a,\theta_0)\right) \left|K_n''(nW_0|a,\theta_0)\right|^{1/2}. \quad (12)$$

Formulas (9) - (12) take a particularly simple form for linear exponential families i.e. for the canonical parameterization $g(\mu_i(\theta)) = \eta_i = \alpha_i + \theta' z_i$. Here, $b_i(a) = 1$ for all i, and from (10)

$$R_n'(t|a,\theta_0) = \sum_{i=1}^{n} z_i (\phi w_i)^{-1} \left[g^{-1}(\alpha_i + \theta_0' z_i + nt' z_i) - g^{-1}(\alpha_i + a' z_i) \right].$$

Suppose that we substitute $W_0 = n^{-1}(a - \theta_0)$ for t; then, the term in the brackets is equal to zero for all i, and thus satisfies the equation $R_n'(W_0|a,\theta_0) = 0$. Substitution of the saddlepoint W_0 into $\hat{\gamma}_i$ yields $\hat{\gamma}_i = \alpha_i + a' z_i$, and therefore that $\dot{c}(\hat{\gamma}_i) = \mu_i(a)$ and $\ddot{c}(\hat{\gamma}_i) = V(\mu_i(a))$. Thus, from (11),

$$K_n''(nW_0|a,\theta_0) = \sum_{i=1}^{n} z_i z_i' \frac{V(\mu_i(a))}{\phi w_i},$$

yielding from (9) and (12) the saddlepoint approximation

$$g_{\hat{\theta}}(a|\theta_0) = \exp\left(\sum_{i=1}^{n} \frac{c(\alpha_i + a' z_i) - c(\alpha_i + \theta_0' z_i) - \mu_i(a)(a - \theta_0)' z_i}{\phi w_i} \right) \left| \sum_{i=1}^{n} z_i z_i' \frac{V(\mu_i(a))}{2\pi \phi w_i} \right|^{1/2}. \quad (13)$$

This approximation may be easily computed for any of the linear exponential families using Table 2.1 of McCullagh and Nelder (1989, page 30) with appropriate changes in notation.

3.2 Exponential Regression

Suppose $Y_i \sim \text{Exp}(\mu_i)$ are independent random variables with means $\mu_i = \theta z_i$ for $\theta > 0$ and nonnegative z_i. Simple calculations lead one to the cumulant generating function for the

score function

$$K_n(t|a,\theta_0) = n\left[\log\theta_0 + ta^{-1} - \log(\theta_0+t)\right].$$

Using the results of Section 3.1 as a guide, the saddlepoint W_0 equals $n^{-1}(a-\theta_0)$ and $K_n''(nW_0|a,\theta_0) = n\theta_0^{-2}$. Thus, by (13),

$$\begin{aligned} g_{\hat\theta}(a|\theta_0) &= \sqrt{\frac{1}{2\pi}}\,\exp\left(K_n(a-\theta_0|a,\theta_0)\right)\,\left(\sum_{i=1}^n z_i^2 \mathrm{var}_a(Y_i)\right)^{1/2} \\ &= \sqrt{\frac{1}{2\pi}}\,\exp\left(n\left(1-\frac{\theta_0}{a}+\log\frac{\theta_0}{a}\right)\right)\,\left(\sum_{i=1}^n z_i^2\frac{1}{a^2 z_i^2}\right)^{1/2} \\ &= \sqrt{\frac{n}{2\pi a^2}}\,\exp\left(n\left(1-\frac{\theta_0}{a}+\log\frac{\theta_0}{a}\right)\right) \end{aligned}$$

is the unnormalized saddlepoint approximation to the density of $\hat\theta$. This approximation is in fact exact up to renormalization and invariant with respect to the values of the covariates. This should come as no surprise since (i) the Y_i can be made independent and identically distributed under a simple reparameterization; and (ii) saddlepoint expansions in the independent and identically distributed setting are known to be exact for the normal, gamma, and inverse gamma families of distributions (Daniels, 1980). In Figure 1, we compare the saddlepoint approximation to the distribution of $\hat\theta$ with the exact density and corresponding normal approximation for $n = 3$. It is equally simple to do such calculations for the more general link function $\mu_i = \exp\{\theta z_i\}$; here, no restrictions on θ or z_i are needed. However, in this case the calculation of the saddlepoint under the non-canonical link function must be done numerically; once found, the formulae in (9)-(12) can be used to obtain the saddlepoint approximation.

3.3 Poisson Regression (single parameter)

Suppose that $Y_i \sim \mathrm{Poisson}(\mu_i(\theta))$, where $\mu_i = \exp(\theta z_i)$. Then,

$$K_n(t|a,\theta_0) = \sum_{i=1}^n \mu_i(\theta_0)\left(\exp(tz_i) - 1\right) - tz_i\mu_i(a),$$

the saddlepoint is once again $W_0 = n^{-1}(a - \theta_0)$, and the approximation to the density of $\hat{\theta}$ is thus given by

$$g_{\hat{\theta}}(a|\theta_0) = \left(\frac{\sum_{i=1}^{n} z_i^2 \exp(az_i)}{2\pi} \right)^{1/2} \exp(K_n(a - \theta_0|a, \theta_0)) .$$

In Figures 2a and 2b, plots of the exact distribution as well as the corresponding saddlepoint and normal approximations to the density of $\hat{\theta}$ are provided for two different covariate patterns and sample sizes. The agreement between all three in Figure 2a ($n = 10$) is evident; however, we see in Figure 2b ($n = 5$) that the normal approximation is inferior. Figure 2c provides a more detailed examination of the left and right tails of the density for the uniformly distributed covariates; the exact density has been estimated using a kernel density estimator (cf. Scott, 1992) applied to 150,000 simulated maximum likelihood estimates. The agreement between the exact distribution and saddlepoint approximation is again excellent. Since the distribution of $\hat{\theta}$ is technically discrete conditionally upon the z_i, a continuity-corrected version of the formula, similar to (5.4) of Daniels (1983), may be used; however, in practice we have found that the difference between the two approximations is negligible.

3.4 Logistic Regression

We shall illustrate the vector parameter approximation using logistic regression with one covariate and an unknown intercept. The cumulant generating function for n independent observations based on such a logistic regression model is

$$K_n(t|a, \theta_0) = \sum_{i=1}^{n} \log [1 - p_i(\theta_0) + p_i(\theta_0) \exp(t'z_i)] - t'z_i p_i(a),$$

where t, a and θ_0 are 2×1 vectors, $p_i(\theta) = [1 + \exp(-(\theta_{00} + \theta_{01}x_i))]^{-1}$ and $z_i = (1, x_i)'$. The two-dimensional saddlepoint $W_0 = n^{-1}(a - \theta_0)$, and the approximate density of $\hat{\theta}$ is given by

$$g_{\hat{\theta}}(a|\theta_0) = \frac{|\sum_{i=1}^{n} z_i z_i' p_i(a)(1 - p_i(a))|^{1/2}}{2\pi} \exp(K_n(a - \theta_0|a, \theta_0)) ,$$

where $|\cdot|$ denotes the determinant. This formula is equally simple for parameter vectors of fixed but arbitrary dimension.

We consider three distinct examples in this section; they have been chosen to correspond to varying degrees of expected difficulty in achieving "goodness-of-approximation". The exact distribution of $\hat{\theta}$ in each case is estimated via smoothed bivariate average shifted histograms (Scott, 1992), and is based on the maximum likelihood estimates obtained from 150,000 simulated datasets. The sample size in each case is $n = 20$, placing us in a situation where the accuracy of the normal approximation to the distribution of $\hat{\theta}$ should be suspect. In each plot, the solid contours represent the simulated exact distribution, and the two dashed contours correspond to the saddlepoint and normal approximations; for the purposes of comparison, each distribution is normalized to have the (highest) mode equal to one, and the contour lines are placed at the same heights for each distribution.

In first example, the covariate chosen is highly discrete, having only four possible distinct values. Hence, the sufficient statistics for the components of $\hat{\beta}$ should share a similar property, resulting in a distribution that is highly discrete. Figure 3a provides the contours for the three distributions; the estimated exact distribution is as expected highly discretized. The saddlepoint approximation, while clearly not able to emulate the highly discrete nature of the density, is centered correctly, sufficiently dispersed, and in fact "lassos" the contours of the exact distribution very well. The normal approximation, while centered correctly, is inadequately dispersed. In the second example, the values of the covariate are randomly generated from the standard normal distribution. As can be seen from Figure 3b, the estimated exact distribution is quite spherical, and the saddlepoint approximation does an excellent job in approximating the density. The normal approximation is centered correctly, but has tails that greatly underestimate the spread of the distribution. In our last example, the covariates have been generated from an exponential distribution having mean equal to one. The contours of the estimated exact distribution, seen in Figure 3c, are quite elliptical,

and the saddlepoint approximation is again extremely accurate; the normal approximation is seen to be woefully inadequate.

In each of the examples presented, it is clear that any inferential procedure based upon the normal approximation will be highly misleading; this is particularly true with regard to observed significance levels, which depend heavily upon the agreement between the tails of the exact and approximate distributions.

4 Implementation

In the univariate setting, calculation of tail probabilities, expectations, or other quantities based on the density estimate may be done via numerical integration. Since the density approximations given thus far are unnormalized, such calculations require the normalizing constant. This can be done rather easily if one has a good numerical integration routine available. Monte Carlo methods such as accept/reject sampling (e.g. Rubinstein, 1981) are ideal in this setting since the normalizing constant is no longer required and a candidate density can often be determined with little work.

In higher dimensional problems, the integration problem becomes less tractable as the dimension of the parameter increases, and one is often forced into using some sort of Monte Carlo method. The method of choice here is often a Metropolis-type algorithm such as the Gibbs sampler. While computationally demanding, these methods are widely applicable, relatively easy to implement, and do not require specification of the normalizing constant. We outline how one might proceed in the multivariate setting in the context of an example. Härdle and Stoker (1989) analyze data from a study on the calibration of "crash dummies" used in automobile safety tests. Data from 58 simulated side-impact collisions are available; the response variable is binary, taking the value one if the collision is judged to have resulted in a fatality. The predictor variables are the age of the subject (reflected in the

design of the crash dummy), the velocity of the automobile, and the maximal acceleration induced on the subjects abdomen at the time of impact. There is evidence in the data that a logistic regression model might not be appropriate, but we shall proceed as if the model were correct. The maximum likelihood estimates and 95% confidence intervals (based on the assumed normality of the MLE) are given in Table 1.

Table 1: Asymptotic Results for Härdle and Stoker Data

Parameter	MLE	95% CI
Intercept	-15.054	(-25.43, -4.68)
Age	0.171	(0.086, 0.256)
Velocity	0.146	(-0.074, 0.366)
Acceleration	0.016	(-0.012, 0.045)

It is easy to obtain the estimated saddlepoint density for the MLE vector using analogous formulae to those in Section 3.4. However, in contrast to the normal approximation, the marginal saddlepoint density for each parameter (obtained by integrating out the remaining variables) depends upon the true values of all of the parameters e.g. the marginal density for $\hat{\beta}_{\text{age}}$ will depend upon β_{age} as well as β_{int}, β_{vel}, and β_{acc}. Thus, simply using numerical integration to marginalize the density is not recommended unless it is known that one has parameter orthogonality. One can get around this problem by taking a Bayesian approach. Specifically, by placing a uniform improper prior on each of β_{int}, β_{age} β_{vel}, and β_{acc}, standard calculations yield that the posterior density of the parameters given the data is proportional to the saddlepoint density of the MLE's. To obtain the marginal posterior density for each parameter, one can first apply the Gibbs sampler to obtain observations from the joint posterior density, and then resort to Monte Carlo marginalization to obtain the desired result. In Figure 4, we provide the estimated marginal posteriors for the data described above. Each curve is based on one long chain of 15,000 sampled observations (using a burn-in of 200); for convenience, we employed the griddy Gibbs sampler on a very fine grid (Ritter and Tanner, 1992). Then, for example, the marginal density of $\beta_{\text{age}}|$ data

was obtained using the formula

$$m_{\beta_{\text{age}}}(u \mid \text{data}) = \frac{1}{K}\sum_{i=1}^{K} \phi\left(\beta_{\text{age}}^{(i)} | \beta_{\text{int}}^{(i)}, \beta_{\text{vel}}^{(i)}, \beta_{\text{acc}}^{(i)}\right) \frac{\pi\left(\beta_{\text{int}}^{(i)}, u, \beta_{\text{vel}}^{(i)}, \beta_{\text{acc}}^{(i)} \mid \text{data}\right)}{\pi\left(\beta_{\text{int}}^{(i)}, \beta_{\text{age}}^{(i)}, \beta_{\text{vel}}^{(i)}, \beta_{\text{acc}}^{(i)} \mid \text{data}\right)},$$

where K is the number of sampled vectors of the form $\left(\beta_{\text{int}}^{(i)}, \beta_{\text{age}}^{(i)}, \beta_{\text{vel}}^{(i)}, \beta_{\text{acc}}^{(i)}\right)'$, $\pi(\cdot \mid \text{data})$ is proportional to the joint posterior (i.e. the unnormalized saddlepoint density), and $\phi(\cdot|\cdot)$ is any proper conditional density function. For simplicity, we let $\phi(\cdot|\cdot)$ be the appropriate (conditional) univariate normal probability density function corresponding to the assumption that $\left(\beta_{\text{int}}, \beta_{\text{age}}, \beta_{\text{vel}}, \beta_{\text{acc}}\right)'$ is multivariate normal with mean given by the MLE and variance given by the sandwich estimator. As can be seen in the plots, the distributions of the age and acceleration parameters are mildly skewed; the distributions of the intercept and velocity parameters are clearly non-normal. The corresponding 95% highest posterior density regions are given in Table 2.

Table 2: Estimated 95% HPD Regions for Härdle and Stoker Data

Parameter	HPD interval
Intercept	(-32.892, -8.856)
Age	(0.112, 0.295)
Velocity	(-0.033, 0.461)
Acceleration	(-0.011, 0.051)

For each variable, it can be seen that the highest posterior density interval is (i) not symmetric about the MLE and (ii) wider than the corresponding 95% confidence interval obtained above. These calculations took approximately 6 hours on a SPARCSTATION 20 using software written by the authors in the languages *S*-plus and FORTRAN. An alternative to using the saddlepoint density to calculate small sample intervals in this example is to use exact methods for logistic regression. The only commercially available software we are aware of that has the capability to perform such calculations is *LogXact Turbo* (Cytel, 1993); unfortunately, it could not solve this problem due to the size of the associated permutation distribution.

References

Bak, J. and Newman, D.J. (1982). *Complex Analysis.* Springer-Verlag, NY.

Barndorff-Nielsen, O.E. and Cox, D.R. (1979). Edgeworth and Saddlepoint Approximations with Statistical Applications. *Journal of the Royal Statistical Society, Series B*, Vol **41**: 279-312.

Barndorff-Nielsen, O.E. and Cox, D.R. (1989). Asymptotic Techniques for Use in Statistics. Chapman-Hall, NY.

Cytel Software Corporation. (1993). *LogXact: Software for Exact Logistic Regression.* Cambridge MA.

Daniels, H.E. (1954). Saddlepoint Approximations in Statistics. *Annals of Mathematical Statistics*, **25**: 631-50.

Daniels, H.E. (1980). Exact Saddlepoint Approximations. *Biometrika*, **67**: 59-63.

Daniels, H.E. (1983). Saddlepoint Approximations for Estimating Equations. *Biometrika*, **70**: 89-96.

Davison, A.C. (1988). Approximate Conditional Inference in Generalized Linear Models. *Journal of the Royal Statistical Society, Series B*, Vol **50**: 445-461.

Durbin, J. (1980). Approximations for densities of sufficient estimators. *Biometrika*, **67**: 311-333.

Estrada, R. and Kanwal, R. P. (1994). *Asymptotic Analysis: A Distributional Approach.* Birkhäuser, Boston.

Field, C.A. (1982). Small sample asymptotic expansions for multivariate M-estimates. *Annals of Statistics*, **10**: 672-689.

Field, C.A. and Ronchetti, E. (1990). Small Sample Asymptotics. Lecture Notes - Monograph Series (Volume 13). *IMS*, Hayward, CA.

Härdle, W. and Stoker, T.M. (1989). Investigating Smooth Multiple Regression by the Method of Average Derivatives. Journal of the American Statistical Association, **Vol 84**: 986-995.

Hinkley, D.V., Reid, N., and Snell, E.J. (eds) (1991). *Statistical Theory and Modeling: In honour of Sir David Cox, FRS.* Chapman Hall.

Hougaard, P. (1985). Saddlepoint Approximations for Curved Exponential Families. *Statistics and Probability Letters,* **Vol 3**: 161-166.

Kolassa, J.E. (1994). Series Approximation Methods in Statistics. Lecture Notes in Statistics. Springer-Verlag, NY.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data.* John Wiley and Sons, NY.

Lugannani, R. and Rice, S. (1980). Saddlepoint Approximation for the Distribution of the Sum of Independent Random Variables. *Adv. Appl. Probab.*, **12**: 475-490.

McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models, 2^{nd} ed. Chapman-Hall.

Ronchetti, E. and Welsh, A.H. (1994). Empirical Saddlepoint Approximations for Multivariate M-Estimators. *Journal of the Royal Statistical Society, Series B*, Vol **56**: 313-326.

Rubinstein, R. (1981). Simulation and the Monte Carlo Method. John Wiley and Sons, NY.

Scott, D. (1992). *Multivariate Density Estimation.* John Wiley and Sons, NY.

Small, C.G. and McLeish, D.L. (1994). Hilbert Space Methods in Probability and Statistical Inference. John Wiley and Sons, NY.

Skovgaard, I.M. (1987). Saddlepoint Expansions for Conditional Distributions. *Journal of Applied Probability,* **Vol 24**, 875-887.

Ritter, C. and Tanner, M.A. (1992). Facilitating the Gibbs Sampler: The Gibbs Stopper and Griddy-Gibbs Sampler. Journal of the American Statistical Association, **Vol 87**: 861-868.

Appendix: Expansions for Sums of Independent Random Variables

Let $X_1 \ldots X_n$ be independent continuous random vectors on $\mathbb{R}^k$, not necessarily identically distributed, and define $S_n = \sum_{i=1}^n X_i$. Let the density function for S_n, say $f_n(s)$, have support on a possibly infinite domain, and define

$$M_n(t) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\{t's\} f_n(s) ds_1 \cdots ds_k,$$

$K_n(t) = \log M_n(t)$, and $\xi_n(s) = M_n(is)$ as the moment generating function, cumulant generating function, and characteristic function respectively. Suppose that the moment generating function exists in an open set containing $t = 0$ and that $\xi_n(s)$ is absolutely integrable. Then,

$$\begin{aligned} f_n(s) &= \left(\frac{n}{2\pi}\right)^k \int_{-i\infty}^{i\infty} \cdots \int_{-i\infty}^{i\infty} M_n(nz)\ \exp\left(-nz's\right)\ dz_1 \cdots dz_k \\ &= \left(\frac{n}{2\pi}\right)^k \int_{-i\infty}^{i\infty} \cdots \int_{-i\infty}^{i\infty} \exp\left(n[R_n(z) - z's]\right)\ dz_1 \cdots dz_k, \end{aligned} \tag{14}$$

where $R_n(z) = n^{-1} K_n(nz)$. By the Closed Curve Theorem (Bak and Newman, 1982, §8.1), the paths of integration include any path from $-i\infty$ to $i\infty$. The last expression results from classical Fourier inversion and a change of variables; for $k = 1$ and independent and identically distributed random variables, this expression is exactly that of Field and Ronchetti (1990, §4.3).

We use the methods of Estrada and Kanwal (1994) to approximate the multivariate complex integral in (14). Specifically, let $h_n(z) = z's - R_n(z)$, and suppose Z_0 is an interior maxima of $h_n(\cdot)$. Then, it follows that

$$\frac{\partial h_n}{\partial z_i}\Big|_{z=Z_0} = 0$$

for $1 \leq i \leq k$, and the Hessian matrix

$$A = \frac{\partial^2 h_n}{\partial z_i \partial z_j}\Big|_{z=Z_0}$$

is positive definite. Applying Morse's Theorem (Milnor, 1963, §2.1), we can find a local change of variables such that $h_n(z) = h_n(Z_0) + |\Psi(z)|^2$, where $\Psi(Z_0) = 0$ and

$$\frac{\partial \Psi(z)}{\partial z_1 \cdots \partial z_k} > 0$$

for z near Z_0. Using the moment expansion in Estrada and Kanwal (1994, §4.3), it follows that

$$\exp\{-nh_n(z)\} = \exp\{-nh_n(Z_0)\} \left[\left(\frac{2\pi}{n}\right)^{k/2} |A|^{-1/2}\, \delta(z - Z_0) + O\left(\frac{1}{n^{k/2+1}}\right)\right], \tag{15}$$

where $\delta(\cdot)$ is the Dirac function. Substitution of (15) in (14) yields

$$f_n(s) = \left(\frac{n}{2\pi}\right)^{k/2} \exp\{-nh_n(Z_0)\}\,|A|^{-1/2}\,i^{-k}\int_{-i\infty}^{i\infty}\cdots\int_{-i\infty}^{i\infty}\delta(z-Z_0)dz_1\cdots dz_k + O\left(\frac{1}{n^{k/2+1}}\right)$$

$$= \left(\frac{n}{2\pi}\right)^{k/2} \exp\left(n[R_n(Z_0) - Z_0's]\right)\;|R_n''(Z_0)|^{-1/2} + O\left(\frac{1}{n^{k/2+1}}\right). \qquad (16)$$

Here, $R_n''(z) = nK_n''(nz)$, and Z_0 solves $R_n'(Z_0) = s$ where $R_n'(z) = K_n'(nz)$. The last step then follows from substituting $R_n(z) - z's$ into the previous expression and evaluating the path integral of the Dirac function, which is equal to one. We note that for $k = 1$ this is exactly formula (1.1) of Easton and Ronchetti (1986), except that it now applies to the sum of independent (but not necessarily identically distributed) random vectors. Based on formula (3.2) of the same paper, it is then easy to derive that

$$\text{pr}\{S_n > a\} \doteq \int_{Z_l}^{\infty}\cdots\int_{Z_k}^{\infty}\left(\frac{n}{2\pi}\right)^{k/2}|R_n''(Z_0)|^{1/2}\;\exp\left(n[R_n(Z_0) - Z_0'R_n'(Z_0)]\right)\;dZ_0 \qquad (17)$$

where Z_l solves $R_n'(Z_l) = a$.

Figure 1: Exponential Regression (n=3)

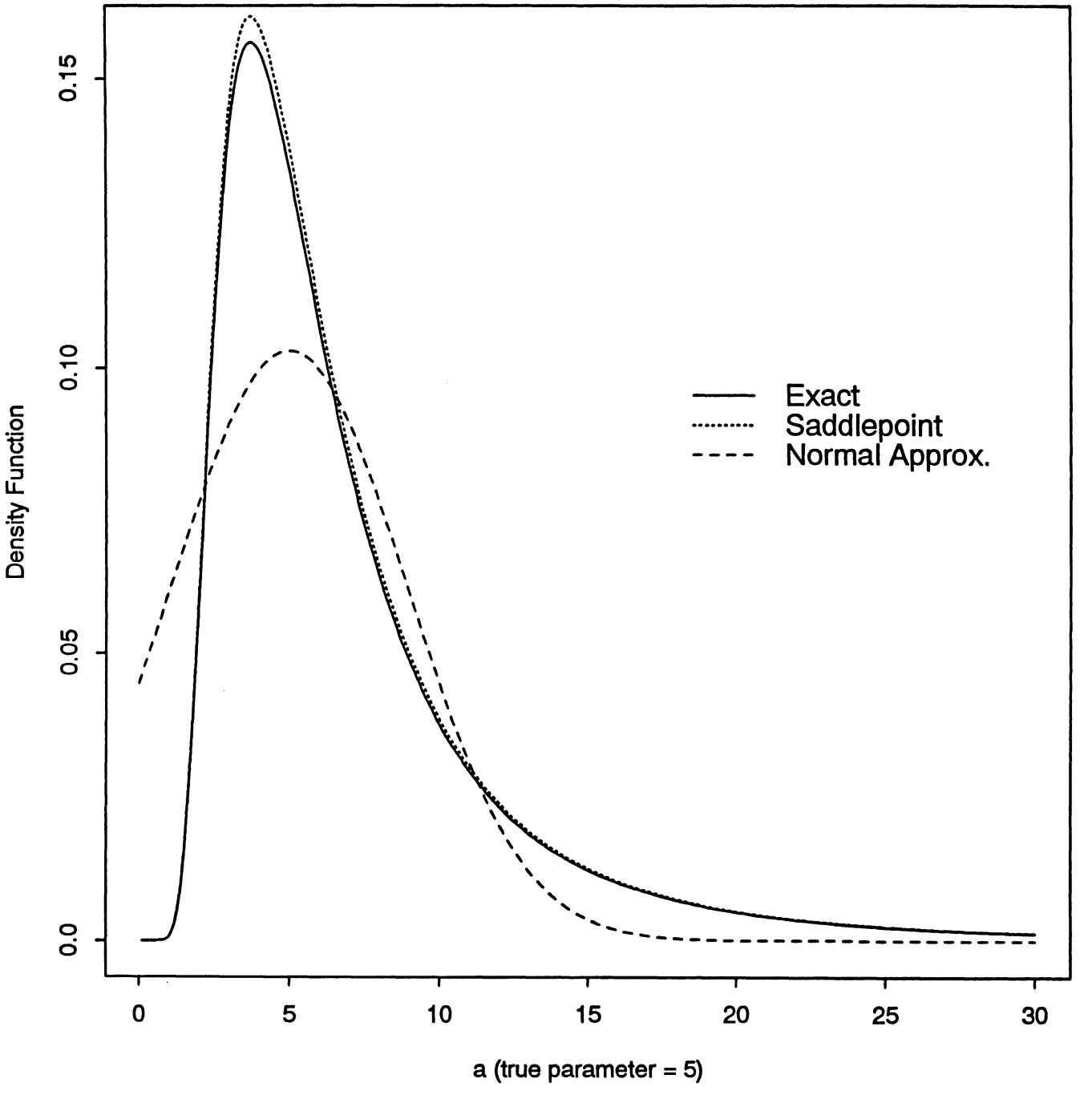

Figure 2a: Poisson Regression (n=10, Z=(1,1,2,2,3,3,4,4,5,5))

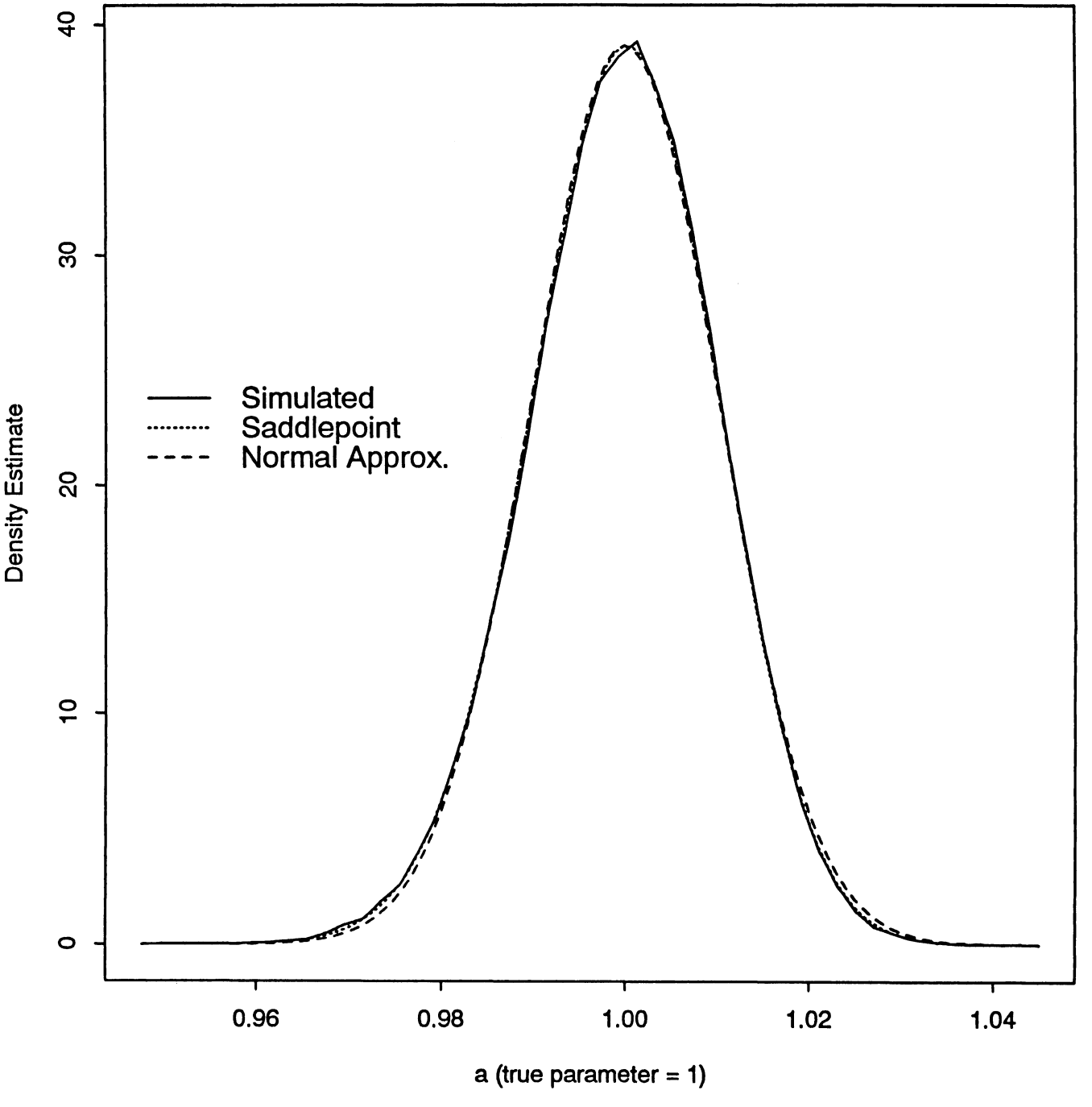

Figure 2b: Poisson Regression (n=5, Z ~ Uniform(0,50))

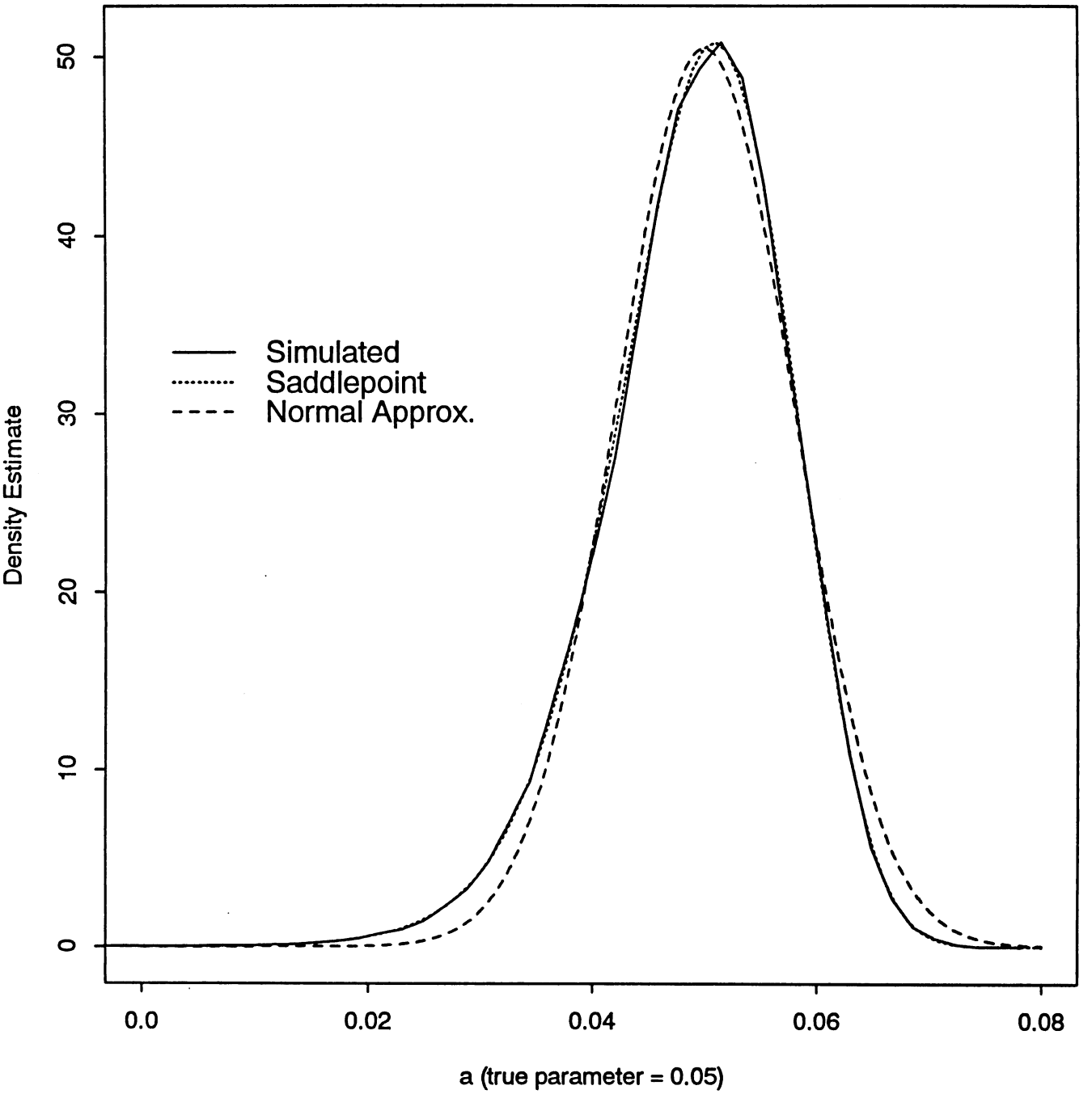

Figure 2c: Poisson Regression (n=5, Z ~ Uniform(0,50))

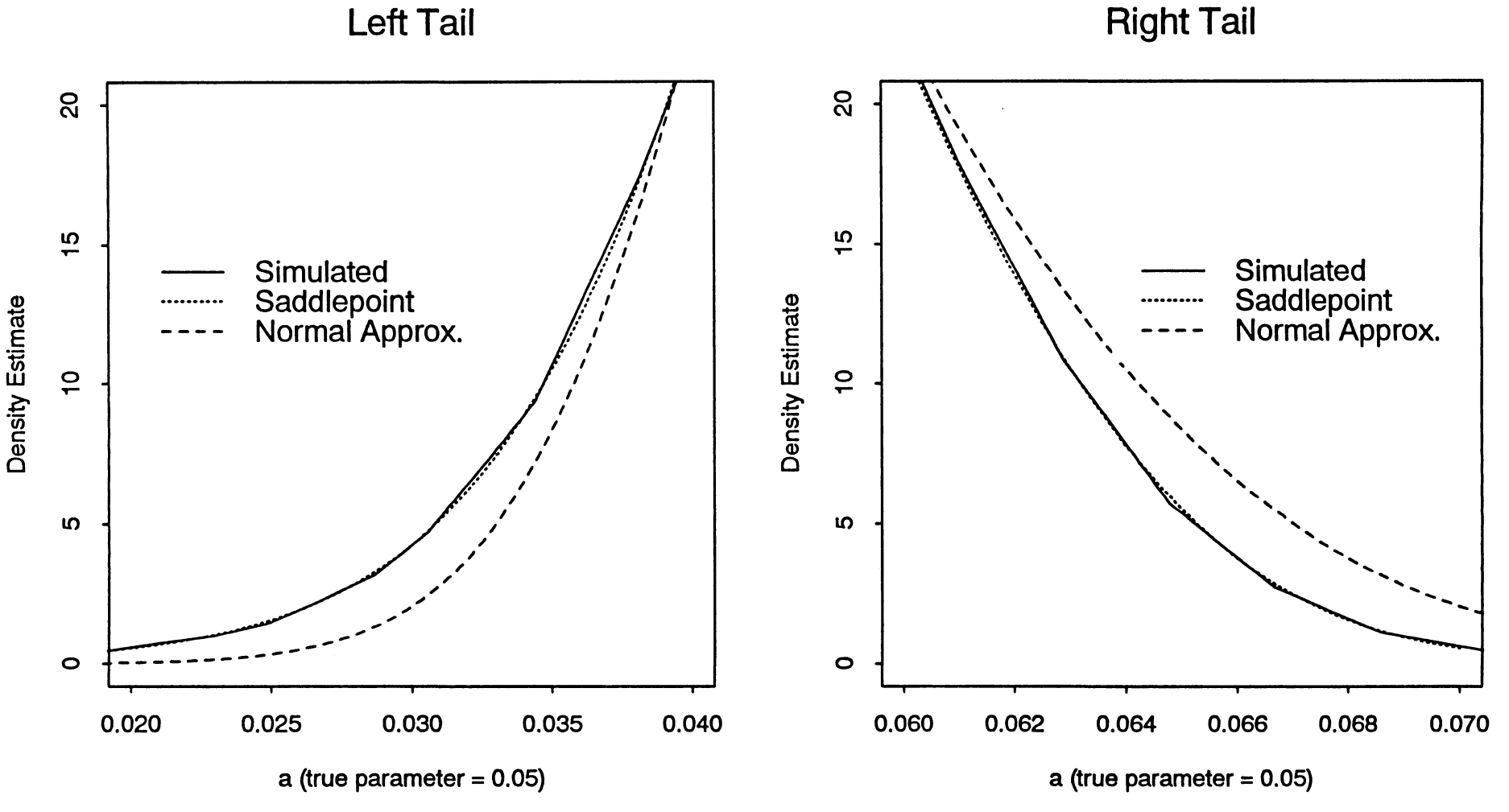

Figure 3a: Distribution Contours for Logistic Regression MLE's (X = (1,2,3,4))

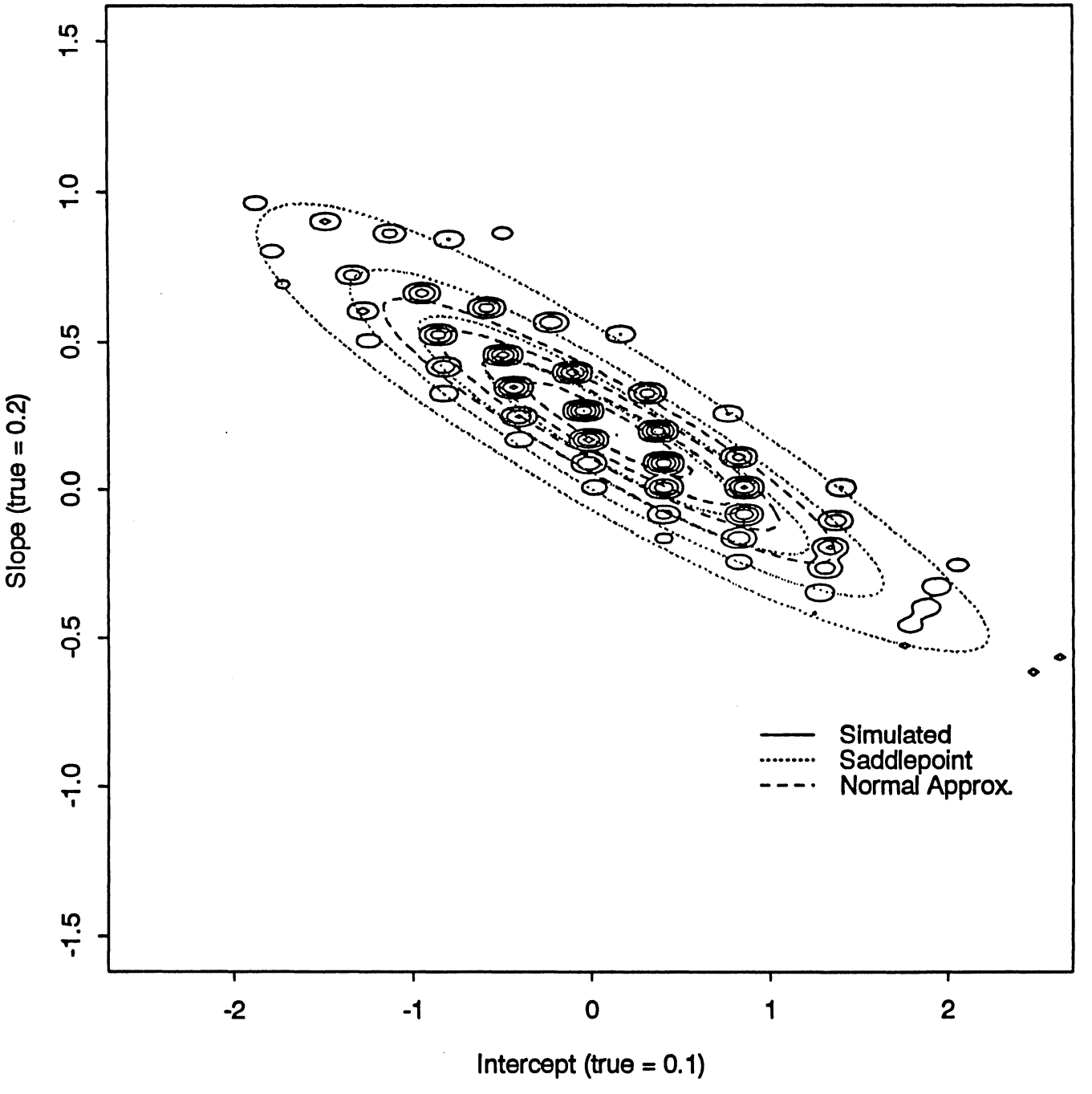

Figure 3b: Distribution Contours for Logistic Regression MLE's (X ~ N(0,1))

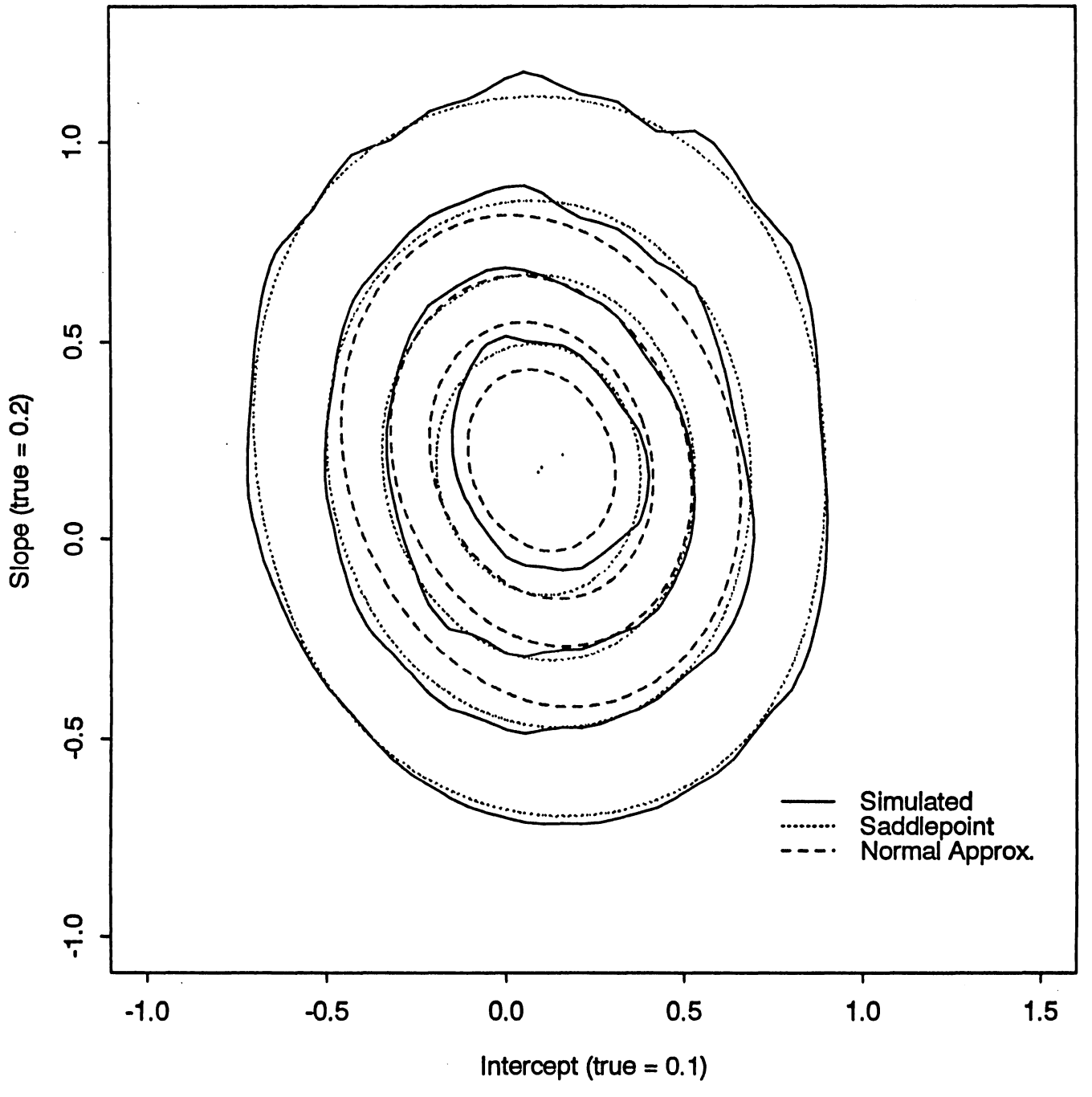

Figure 3c: Distribution Contours for Logistic Regression MLE's (X ~ Exp(1))

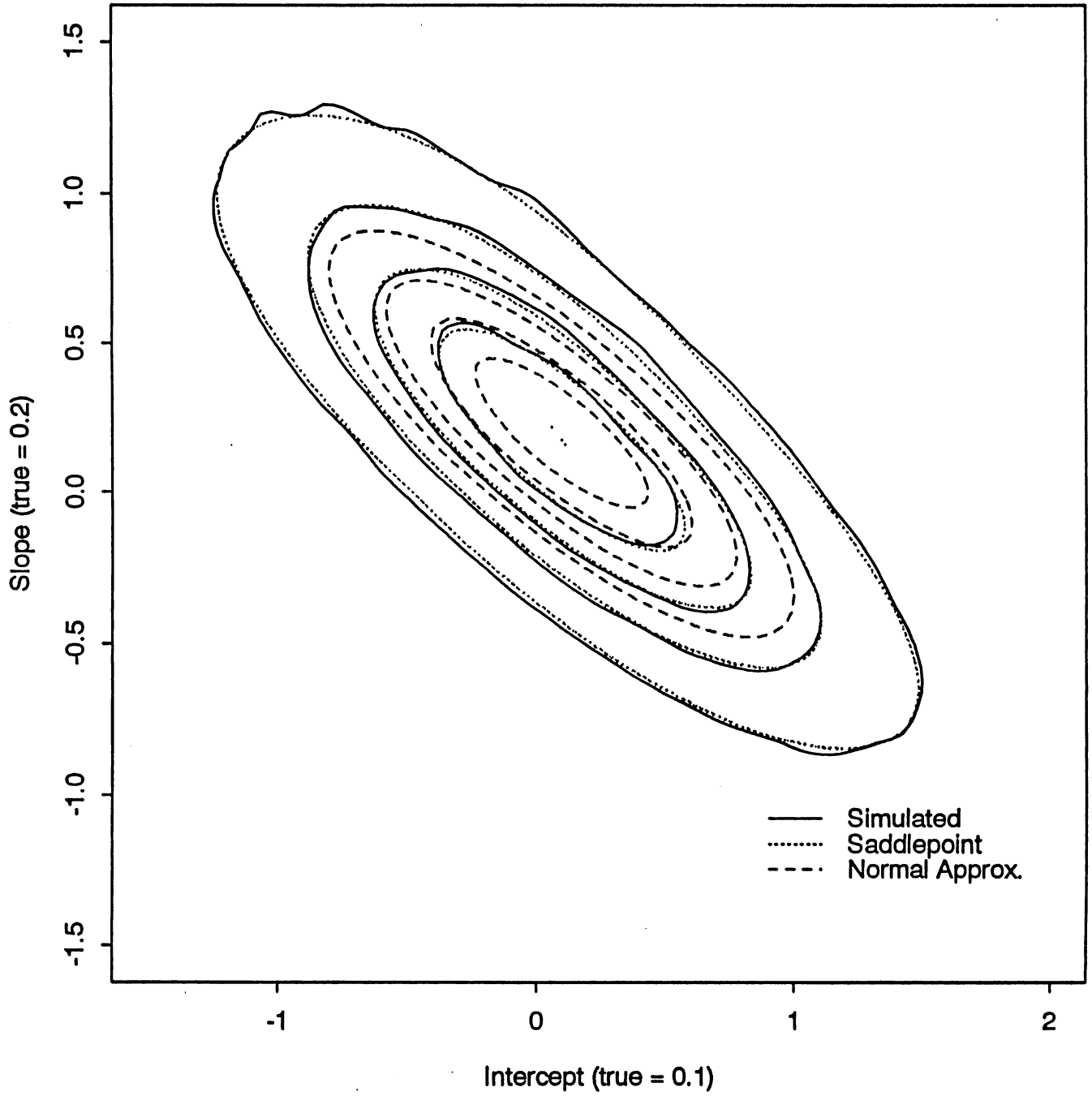

Figure 4: Marginal Posteriors of Regression Parameters